BREA5292-R1

A gene expression signature identifies two prognostic subgroups of basal breast cancer

Renaud SABATIER[1,2], Pascal FINETTI[1], Nathalie CERVERA[1],
Eric LAMBAUDIE[3], Benjamin ESTERNI[4], Emilie MAMESSIER[5],
Agnès TALLET[6], Christian CHABANNON[7,8], Jean-Marc EXTRA[2],
Jocelyne JACQUEMIER[9], Patrice VIENS[2,8],
Daniel BIRNBAUM[1], François BERTUCCI[1,2,8]

Short title : Multigene prognostic classification of basal breast cancers

Addresses

1) Département d'Oncologie Moléculaire, Centre de Recherche en Cancérologie de Marseille, UMR891 Inserm, Institut Paoli-Calmettes, Marseille, France

2) Département d'Oncologie Médicale, Institut Paoli-Calmettes, Centre de Recherche en Cancérologie de Marseille, UMR891 Inserm, Marseille, France

3) Département de Chirurgie, Institut Paoli-Calmettes, Centre de Recherche en Cancérologie de Marseille, UMR891 Inserm, Marseille, France

4) Bureau d'Etudes Cliniques, Institut Paoli-Calmettes, Centre de Recherche en Cancérologie de Marseille, UMR891 Inserm, Marseille, France

5) Centre d'Immunologie de Marseille-Luminy, Marseille, France

6) Département de Radiothérapie, Institut Paoli-Calmettes, Centre de Recherche en Cancérologie de Marseille, UMR891 Inserm, Marseille, France

7) Centre de Ressources Biologiques, Institut Paoli-Calmettes, Centre de Recherche en Cancérologie de Marseille, UMR891 Inserm, Marseille, France

8) Université de la Méditerranée, Marseille, France

9) Département d'Anatomopathologie, Institut Paoli-Calmettes, Centre de Recherche en Cancérologie de Marseille, UMR891 Inserm, Marseille, France.

Corresponding author: Prof François Bertucci. at: Département d'Oncologie Médicale, Institut Paoli Calmettes, UMR891 Inserm, 232 Bd. Sainte-Marguerite, 13273 Marseille Cedex 09, France.

Tel: 33 4 91 22 35 37; Fax : 33 4 91 22 36 70;

E-mail: bertuccif@marseille.fnclcc.fr

ABSTRACT

Prognosis of basal breast cancers is poor but heterogeneous. Medullary breast cancers (MBC) display a basal profile, but a favorable prognosis. We hypothesized that a previously published 368-gene expression signature associated with MBC might serve to define a prognostic classifier in basal cancers. We collected public gene expression and histoclinical data of 2145 invasive early breast adenocarcinomas. We developed a Support Vector Machine (SVM) classifier based on this 368-gene list in a learning set, and tested its predictive performances in an independent validation set. Then, we assessed its prognostic value and that of six prognostic signatures for disease-free survival (DFS) in the remaining 2034 samples. The SVM model accurately classified all MBC samples in the learning and validation sets. A total of 466 cases were basal across other sets. The SVM classifier separated them into two subgroups, subgroup 1 (resembling MBC) and subgroup 2 (not resembling MBC). Subgroup 1 exhibited 71% 5-year DFS, whereas subgroup 2 exhibited 50% (p=9.93E-05). The classifier outperformed the classical prognostic variables in multivariate analysis, conferring lesser risk for relapse in subgroup 1 (HR=0.52, p=3.9E-04). This prognostic value was specific to the basal subtype, in which none of the other prognostic signatures was informative. Ontology analysis revealed effective immune response, enhanced tumor cell apoptosis, elevated levels of metastasis-inhibiting factors and low levels of metastasis-promoting factors in the good-prognosis subgroup, and a more developed cell migration system in the poor-prognosis subgroup. In conclusion, based on this 368-gene SVM model derived from an MBC signature, basal breast cancers were classified in two prognostic subgroups, suggesting that MBC and basal breast cancers share similar molecular alterations

associated with aggressiveness. This signature could help define the prognosis, adapt the systemic treatment, and identify new therapeutic targets.

Key words

Basal breast cancer, DNA microarrays, medullary, prognosis.

List of abbreviations

DFS: disease-free survival; DWD: Distance Weighted Discrimination; ER: estrogen receptor; GEO: Gene Expression Omnibus; HR: hazard ratio; IHC: immunohistochemistry; IPC: Institut Paoli-Calmettes; MBC: medullary breast cancer; NCBI: National Cancer for Biotechnology Information; PR: progesterone receptor; SBR: Scarff Bloom Richardson; SSP: Single Sample Predictor; SVM, Support Vector Machine.

INTRODUCTION

Prognosis of breast cancer is heterogeneous and imperfectly captured by classical histoclinical features, making clinical evolution difficult to predict for a given patient and treatment not perfectly adapted. Over the past decade [1], gene expression profiling revealed five molecular subtypes of breast cancer based on the expression patterns of an intrinsic gene set: luminal A and B, basal, ERBB2 and normal-like [2]. These subtypes represent different disease entities associated with specific molecular alterations and histoclinical features [3-7]. This classification correlates with major prognostic variables. Thus, its added prognostic value remains unclear. However, it provides the opportunity to investigate biological questions, such as the identification of prognostic or therapeutic targets, in more homogenous entities, and therefore enrich for a signal relevant in a specific subtype, which would be diluted and undetectable in the whole breast cancer population [8]. For example, the predictive value of *P53* mutations regarding the response to chemotherapy is opposite according to luminal A or basal subtype [9]. In a recent meta-analysis [10], similar observations were done with seven prognostic multigene expression signatures [11-18], which were highly informative of clinical outcome in ER+/ERBB2- cases, but much less informative and never significant in ER-/ERBB2- and ERBB2+ cases. One of the reason is that three of the four signatures defined by supervised analysis had been initially defined in ER+ tumors [13,15], or by separately analyzing ER+ and ER- tumors but with a few ER- cases and without taking into account the heterogeneity of ER- tumors [17]. Few studies have attempted to derive prognostic signatures of ER- breast cancer [19-22], and all showed the difficulty of the task.

The basal subtype represents around 15% of invasive breast cancers. Most basal breast cancers are of ductal type [23]. Despite a relative chemosensitivity when

compared to other subtypes they display a poor prognosis after treatment, which generally includes adjuvant chemotherapy. However, this subtype shows prognostic heterogeneity since not all basal breast cancer patients have an unfavorable outcome. To date, reliable identification of basal breast cancer patients with a good or a poor prognosis is difficult and based only on histological features thus far from optimal [24,25]. If optimized, this would help tailor treatment by using more or less aggressive approaches based on the prediction of outcome, all the more so as different types of chemotherapy are available, and promising targeted molecular therapies are under development for these tumors [26,27]. Furthermore, such optimization, if based on molecular data, should help identify new potential therapeutic targets.

Medullary breast cancers (MBC) represent less than 2% of breast cancers. Despite features of aggressiveness and the fact that they frequently display a basal profile [28,29], MBC are associated with a favorable prognosis. Using whole-genome oligonucleotide microarrays, we recently reported a list of 368 genes differentially expressed between basal MBC and nonMBC. Here, we used MBC as a model of good-prognosis basal breast cancer and tested the hypothesis that this gene signature might be used to derive a gene classifier predictive for disease-free survival (DFS) in a large pooled data set of basal breast cancers.

MATERIALS AND METHODS

Tumor samples

We collected personal and public data from breast cancer samples profiled using DNA microarrays. Inclusion criteria included: pre-treatment sample of an invasive adenocarcinoma, non-inflammatory and non-metastatic, with available histoclinical data, and profiled using Affymetrix or Agilent oligonucleotide microarrays. All data sets were retrospective. They are described in Supplementary Table 1. The Weigelt's set [30] was used to validate the MBC-nonMBC SVM classifier. Other sets (thereafter designated "prognostic series") were pooled to test the prognostic impact of this classifier in basal breast cancers and other subtypes; they included our series (Institut Paoli-Calmettes – IPC) and 15 public series. The IPC series contained frozen tumor samples obtained from 266 early breast cancer patients who underwent initial surgery in our institution between 1992 and 2004. They included 227 cases previously reported [31] and 39 additional cases, all similarly profiled using Affymetrix U133 Plus 2.0 human oligonucleotide microarrays as previously described [31]. The study was approved by the IPC review board, and informed consent was available for each case. This IPC series did not include the 37 cases from which we had derived the 368-gene list. The 15 public series originated from 15 publications [11-13,17,32-42]. Overall, data from a total of 3409 patient's samples were collected. When different publications included the same patients, the redundancy was eliminated, resulting in 2145 different patient's samples (111 in the Weigelt's set and 2034 in the "prognostic series"). Gene expression and histoclinical data of public series were retrieved from NCBI GEO databases and authors' website. Histoclinical data of our IPC series are available in Supplementary Table 2.

Gene expression data pre-processing

Before analysis, we mapped hybridization probes across the two technological oligonucleotide-based platforms (Agilent and Affymetrix) used across the series. Affymetrix gene chips annotations were updated using NetAffx Annotation files (www.affymetrix.com; release from 01/12/2008). Agilent gene chips annotations were retrieved and updated using both SOURCE (http://smd.stanford.edu/cgi-bin/source/sourceSearch) and EntrezGene (Homo sapiens gene information db, release from 09/12/2008, ftp://ftp.ncbi.nlm.nih.gov/gene/). All probes were thus mapped based on their EntrezGeneID. When multiple probes were mapped to the same GeneID, the one with the highest variance in a particular dataset was selected to represent the GeneID.

Data sets were then processed as follows. For the Agilent-based sets, we applied quantile normalization to available processed data. Regarding the Affymetrix-based data sets, we used Robust Multichip Average (RMA) [43] with the non-parametric quantile algorithm as normalization parameter. RMA was applied to the raw data from the other series and the IPC series. Quantile normalization or RMA was done in R using Bioconductor and associated packages.

Gene expression data analysis

Analysis of each processed data set was done separately to guarantee a larger number of genes common with the intrinsic gene set, the 368-gene list, and the published prognostic signatures.

The molecular subtypes related to the intrinsic breast cancer classification were determined using Hu's Single Sample Predictor (SSP) classifier based on a list of 306 intrinsic genes [35]. We first identified the genes common between the intrinsic gene set and each expression data set. We then used Distance Weighted

Discrimination (DWD) [44] to normalize each data set in order to be comparable to the 315 samples of the Hu's combined test sample set. Next, we defined the expression centroid of each subtype for the common probe sets in this combined test sample set [35]. Finally, we measured the correlation of each sample with each centroid. The sample was attributed the subtype corresponding to the most correlated centroid.

Before constructing our classifier based on the 368-gene list and testing its performances, as well as those of other prognostic signatures, in several independent data sets, gene expression levels of each data set were standardized using the luminal A population as reference, thus allowing to make comparable all data sets. In a previous study [28], we had identified 534 probe sets differentially expressed between basal MBC and basal nonMBC. They represented 368 unique genes. We hypothesized that some nonMBC might have an expression profile close to that of MBC for these genes, and perhaps better prognosis than other nonMBC. Based on this 368-gene list (Supplementary Table 3), we defined a genomic classifier using Support Vector Machine (SVM). The initial outcome of interest for this SVM classifier was the separation MBC-nonMBC, with the secondary objective, in the case where all nonMBC would not be correctly classified, to compare the clinical outcome of these two nonMBC subgroups. Briefly, the SVM algorithm finds in the learning set the hyperplane separating the two subgroups with maximized margins in Euclidean space. This approach allows separation by mapping the data into a high-dimensional feature space that can be linearly separated via a kernel function. Once trained, the classification only requires the linear computation using the learned hyperplane equation. The resulting SVM classification score defines the distance from the hyperplane and so the membership to one of the two subgroups First, we established

the SVM model in our original 37-sample data set (learning set: 16 basal nonMBC and 21 basal MBC) [28], with a polynomial kernel of degree 3. This model classified samples in two subgroups (1 and 2). To test the stability of the SVM model according to the composition of the learning set, we applied 100 random subsamplings to the 37 sample-set by splitting it into training (two thirds of samples) and validation data (remaining one third). For each iteration, the predictive accuracy of the model fitted to the training data was assessed using the validation data, and the results were then averaged over the 100 splits. After having identified the genes common between the 368-gene list and each data set, the SVM model was applied within each set, notably the Weigelt's independent validation set, to attribute each sample to subgroup 1 or 2.

For comparison with our classifier, we tested the predictive value of six major recently reported prognostic breast cancer signatures: the 70-gene signature [11,12], the 76-gene signature [17], the invasiveness gene signature [18], the wound response signature [14], the genomic grade index [13], and the 21-gene recurrence score [15]. Each signature was applied to each expression data set separately to compute a relapse risk to each sample. We first identified the genes common between the signature and each data set. We then strictly applied the same methodology (score or Pearson correlation, and scaling methods) as that reported in the original publications to classify each sample. The original cut-off then defined the membership to the predicted good-prognosis group or the predicted poor-prognosis group. More details are available in Supplementary Table 4.

The 368-gene list was interrogated using Ingenuity software (Redwood City, CA, USA) to assess significant representation of biological pathways

Statistical analysis

Correlations between sample groups and histoclinical factors were calculated with the Fisher's exact test for qualitative variables with discrete categories, and the Mann-Whitney test for continuous variables. Disease-free survival (DFS) was calculated from the date of diagnosis until date of first relapse whatever its location (local, regional or distant) or date of death (when the relapse data was not available) using the Kaplan-Meier method. Survival was compared between groups with the log-rank test. Follow-up was measured from the date of diagnosis to the date of last news for patients without any event. In the "prognostic series" (2034 patients), follow-up was available for 1752 patients (median 88 months) and the 5-year DFS was 68%. Univariate and multivariate analyses were done using Cox regression analysis. The variables tested in univariate analyses included age of patients, pathological tumor size, axillary lymph node status, and SBR grade, ER, PR, ERBB2, and Ki67 IHC status, our SVM model and several published prognostic multigene signatures. Multivariate analysis was done by incorporating all variables with a p-value inferior to 0.05 in univariate analysis. The p-values were based on the Wald test, and patients with one or more missing data regarding the retained variables were excluded. All statistical tests were two-sided at the 5% level of significance. Statistical analysis was done using the survival package (version 2.30), in the R software (version 2.9.1).

RESULTS

Intrinsic molecular subtypes

We collected publicly available gene expression and histoclinical data of a total of 2145 distinct invasive breast adenocarcinomas. We determined the molecular

subtype of tumors in each data set separately by using the SSP method [35]. The percent of genes common to each set and the intrinsic 306-gene set ranged from 88 to 100%.

In the Weigelt's series [30], which contained 10 MBC and 101 non-MBC, 32 tumors were determined as basal. As previously reported by us and others [28,29], all 10 MBC were basal subtype. In this series, only the pathological type was available.

In the "prognostic series" (2034 samples), available histoclinical data allowed us to verify the coherence of the subtypes, notably the basal subtype, in term of histoclinical correlations (Table 1). As expected, basal tumors (466 cases) were diagnosed in younger patients when compared with luminal A tumors. They were also more frequently ductal, with higher pathological size and grade; they were more frequently negative for ER and PR as defined by immunohistochemistry (IHC), but more often positive for P53 and Ki67. Clinical outcome, available for 1752 patients, strongly correlated with subtypes with 5-year DFS of 80% for luminal A (149 events), 57% for luminal B (152 events), 72% for normal-like (78 events), 60% for basal (157 events), and 60% for ERBB2+ (107 events). These results, obtained in a large series of samples, confirmed previous observations, and the coherence of our data set.

The 368-gene model defines two subgroups of basal breast cancers

We used the 37 IPC samples (16 nonMBC and 21 MBC) from which we had generated the 368-gene signature to construct a SVM classifier. As expected, its application onto this learning set resulted in a correct classification of all MBC in subgroup 1 and 14 out of 16 nonMBC in subgroup 2 (Figure 1A). By cross-validation using 100 random iterations, the mean predictive accuracy was 73% (95%CI, [69 –

76]), suggesting the robustness of our SVM model. Its robustness was further tested in an independent public data set [30], which included 10 MBC and 101 nonMBC, of which 32 were basal. The SVM model applied to these 32 tumors (Figure 1B) correctly classified all 10 MBC samples in subgroup 1, whereas 10 out of 22 nonMBC were classified in subgroup 2 and 12 in subgroup 1. This observation suggested that the model is very sensitive for MBC prediction, and that basal nonMBC are heterogeneous regarding the model, with some cases (subgroup 1) resembling more than other cases (subgroup 2) to MBC. This confirmed our initial hypothesis and allowed us to compare the prognosis of these two nonMBC subgroups.

Finally, to test for a prognostic value of such subgrouping, the SVM classifier was applied to each public series separately. The percent of genes common between the 368-gene list and each data set ranged from 72 to 100%. In each set, two tumor subgroups were obtained. Among the 466 pooled basal samples, 217 samples were in subgroup 1 and 249 in subgroup 2 (Figure 2A).

Histoclinical features and prognosis of the two basal subgroups

We compared the histoclinical features of the two subgroups of basal tumors defined by the SVM classifier (Table 2). Out of the nine tested variables, differences (Fisher's exact test) were observed only for age of patients and ER status, with younger patients (p=4.95E-03) and more ER- cases (p=9.14E-04) in subgroup 1. There was no significant difference regarding pathological tumor size, axillary lymph node status, grade, and IHC status for PR, ERBB2, P53 and KI67. Survival information was available for 392 of 466 patients. The survival curves are shown in Figure 2B. With a median follow-up of 81 months, 5-year DFS was better - despite a longer follow-up - for subgroup 1 patients (71% DFS) than for subgroup 2 patients

(50% DFS, p=9.93E-05, log-rank test). Analysis by data set showed that the mean difference of 5-year DFS between subgroups 1 and 2 was 17% (95%CI, [4 – 30], p=0.016). For comparison, the 5-year DFS of our 21 basal MBC [28] was 89%. Thus, our 368-gene classifier identified within basal tumors two subgroups with different prognosis. Subgroup 1 was associated with relatively good prognosis, close to that of normal-like subtype.

Because the prognosis of basal breast cancer is usually unfavorable, most patients are treated using adjuvant chemotherapy. To determine more precisely the link of our predictor with metastatic risk and/or with chemosensitivity, we analyzed patients' subgroups based upon the systemic treatment they received. A total of 116 out of the 392 basal breast cancer patients with available follow-up had not received any adjuvant chemotherapy and hormonal therapy after surgery. With a median follow-up of 102 months, the 5-year DFS was 68% in subgroup 1 and 44% in subgroup 2 (p=7E-03, log-rank test). Thirty basal samples were available in the Hess' series treated with primary chemotherapy before surgery, allowing assessment of response to chemotherapy. The pathological complete response (pCR) rate was 58% in subgroup 1 *vs* only 17% in subgroup 2 (not significant). Altogether, these observations suggested that our genomic predictor is at least associated with prognosis in term of relapse risk, whereas its likely association with response to chemotherapy needs to be tested in larger series.

Univariate and multivariate analyses

We performed univariate and multivariate Cox survival analysis in the combined data set of 466 basal tumors to compare the prognostic performance of the 368-gene classifier (subgroups 1 and 2) with that of histoclinical variables. These

included age of patients, pathological tumor size, axillary lymph node status, SBR grade, and ER, PR, ERBB2, and Ki67 IHC status (Table 3A). In univariate analysis, the HR for relapse was 0.53 for subgroup 1 basal tumors compared to subgroup 2 tumors (95%CI [0.38 – 0.73], p=1.30E-04). Positive lymph node status was associated with DFS whereas age, pathological tumor size, grade, and ER, PR ERBB2, and Ki67 status were not.

Multivariate analysis was done on the 343 out of 466 samples with available information regarding the two significant variables in univariate analysis (lymph node status and 368-gene classifier). Both remained significant (Table 3A), suggesting their independent prognostic value. The multigene classifier was the most significant, with a HR for relapse of 0.52 for subgroup 1 basal tumors compared to subgroup 2 tumors (95%CI [0.36 – 0.74], p=3.9E-04), suggesting a higher prognostic value than the lymph node status (pN). This gene classifier added prognostic information when combined with pN. Indeed, a clinico-genomic model combining pN and the gene classifier performed better than pN alone regarding the prediction of DFS (p=1.7E-4 *vs* p=9.5E-3 respectively, Wald test). By contrast, predictive performances were very close for the clinico-genomic model and the gene classifier alone (p=1.7E-4 *vs* p=1.2E-4 respectively, Wald test).

Comparison of our model with two immune response signatures

Two prognostic immune signatures have been reported in ER- breast cancer: the 7-gene immune response (IR) signature [20,21] and the T-cell metagene [19]. Since our model was enriched with immune response genes (see below), we evaluated its correlation with these two signatures. No gene overlapped our model and the IR signature. Comparison of the classifications of the 466 basal samples

based upon our SVM model and the IR signature showed concordance (both subgroup 1 and IR good-prognosis, or both subgroup 2 and IR poor-prognosis) for only 249 samples (53%). The same analysis was done with the T-cell metagene (50 genes) and our SVM model, revealing only 2 overlapping genes and 66% of concordant classification.

For further comparison, we repeated the prognostic analysis by incorporating the classifications based on these two immune signatures (Table 3B). The 7-gene immune response signature was not significant in univariate analysis, whereas the T-cell metagene was. In multivariate analysis incorporating the three variables significant in univariate analysis, the 368-gene classifier was still the strongest predictor of DFS, independently of the lymph node status, whereas the T-cell metagene lost its prognostic value.

These observations confirm that our model, which includes immune genes and many others, is different from these immune signatures. However, they still highlight the role of immunity in the prognosis of basal breast cancer. To try to elucidate the type of immune response observed in good-prognosis subgroup 1, we searched for correlations between our classification and any of the cell-type specific gene expression profiles of leukocytes. Therefore, we determined if the genes that belong to publicly available B-cell, T-cell, CD8+ T-cell, lymphocyte, and granulocyte signatures [45] were overrepresented in the list of differentially expressed genes between subgroup 1 and subgroup 2 using our 75-basal sample IPC data set. Using the gene set enrichment analysis (GSEA) algorithm [46] and 1.000 permutations, we obtained significance for the B-cell, T-cell, and CD8+ T-cell signatures (Supplementary Figure 1).

The prognostic impact of the 368-gene classifier is specific to the basal subtype

We investigated whether our 368-gene classifier had a prognostic value in the whole population of tumors and in the other subtypes across the “prognostic series” of 2034 samples. The results are summarized in Table 4. When tested in the whole population (1752 patients with annotated follow-up), the 368-gene classifier did not retain any prognostic value. For the analysis per subtype, survival information was available in 243 patients with ERBB2 tumor, 578 patients with luminal A tumor, 304 with luminal B tumor, and 235 with normal-like tumor, with a median follow-up ranging from 77 to 95 months according to the subtype. In each non basal subtype, survival differences between subgroups 1 and 2 were not significant. These results revealed the prognostic specificity of the 368-gene classifier, limited to the basal subtype.

Absence of prognostic impact of other prognostic breast cancer signatures in the basal subtype

We assessed the predictive value of six major prognostic expression signatures recently reported in early breast cancer. In each data set, each sample was assigned a good or a poor prognosis based on each signature. Data sets were then pooled and survival was compared between the predicted good-prognosis and poor-prognosis subgroups. Results of univariate analysis are shown in Table 4. In the whole population, each signature-based classification was strongly associated with DFS, further confirming their robustness. However, when the same analysis was done per molecular subtype none of the signatures retained any prognostic value in the basal tumors, whereas one remained associated with survival in the ERBB2

subtype, four in the luminal B and normal-like subtypes, and six in the luminal A subtype. These results show the subtype-dependence of these signatures regarding their prognostic value, and notably the absence of informative value in the basal subtype, by contrast with our classifier.

DISCUSSION

The extent of the differences between the molecular subtypes evidenced by high-throughput gene profiling makes it necessary to redefine prognostic and predictive markers in each subtype [8]. The interest in such an approach has already been evidenced by the fact that several prognostic multigene signatures [12-18] are highly informative regarding the prediction of clinical outcome in ER+/ERBB2- tumors, but much less in other tumors. Another recent meta-analysis confirmed the interest of the analysis per subtype for developing a more accurate prognostic signature [47]. The rare studies reported for ER- tumors show the difficulty of identifying prognostic multigene signatures in this group [19-22]. This could be because ER- tumors include both basal and ERBB2 samples although these two subtypes are different at the molecular level. No current signature is associated with basal tumors. To our knowledge, the present study is the first one that focuses specifically on basal breast cancers.

We demonstrate that basal breast cancers can be divided into two prognostic subgroups based on a 368-gene expression classifier associated with MBC, a rare histological type associated with a basal profile but a good prognosis. Subgroup 1 exhibited greater DFS (71 *vs* 50%) than subgroup 2, with a HR for relapse was 0.52. Although DFS of subgroup 1 remains insufficient - and cannot preclude the use of

adjuvant systemic therapy - these differences suggest that subgroup 2 patients should need a more aggressive treatment than subgroup 1 patients. Clinico-pathological differences between both subgroups were slight. Like MBC, subgroup 1 tumors were more frequently ER- than subgroup 2 tumors. However, they were not different with respect to the other histological variables, except for age, younger in subgroup 1. This is one of the aims of genomics: to identify molecular subgroups with prognostic relevance within tumors similar at the clinico-pathological level but different regarding their clinical outcome. Multivariate analysis showed that the classifier outperformed the classical histoclinical variables. Its prognostic value was specific to basal tumors. By contrast, none of six major prognostic breast cancer signatures was associated with DFS in these tumors, whereas all were strongly significant in the whole population, and in the luminal A good-prognosis molecular subtype. This observation reinforces the value of our 368-gene classifier in the basal tumors as well as the interest of the per subtype approach.

The strength of our results lies also in the original “bottom-up” approach that we have applied. Unlike the classical “top-down” supervised approach [48], the “bottom-up” approach is based on a starting hypothesis. It consists in first identifying, often in a relatively reduced series, a signature associated with a phenotypic feature relevant to the relapse process (MBC *vs* nonMBC here), and then subsequently testing for its correlation with outcome in a large and independent series of samples, avoiding the problem of overfitting. In the past, prognostic signatures associated with wound repair [49], stem cells [18], P53 mutations [16], pathological grade [13], or inflammatory breast cancer status [50] have been established this way, thereby linking these concepts to tumor cell aggressiveness. Here, we used MBC as a model of non-aggressive basal breast cancer, and showed that the associated signature

holds the fingerprints of the relative more favorable prognosis of this histological type, and that the underlying biological bases are also relevant in basal breast cancers in general.

Since our two basal subgroups were very similar at the clinico-pathological level, specific molecular differences should reside in the 368-gene classifier. What genes and functions represented in this classifier might be associated with a more or less aggressive phenotype is a crucial issue as their identification will help develop targeted therapeutic strategies. Ontology analysis revealed several potentially interesting pathways. Using the Onto-Express algorithm we had previously shown [28] that “Immune response” was the biological process the most represented among the 202 signature genes overexpressed in the good-prognosis subgroup 1 tumors, followed by processes related to apoptosis, and proteolysis. Even if speculative, it is tempting to associate these processes with the better prognosis observed. Ingenuity analysis of canonical pathways confirmed these results (Supplementary Table 5). Our previous analysis had revealed the implication of T_H1 cells in immune response, and a likely high global cytotoxic activity. Two Ingenuity pathways were associated with IL-15 and IL-12. IL-15 is a critical factor for the proliferation and activation of NK and CD8+ T cells. IL-12, and more recently IL-15 and IL-27, demonstrated anti-tumor activity in murine models [51]. The presence of *NFkB2* and *NFkBIE*, an inhibitor of NFkB, in pathways respectively associated with subgroup 1 and subgroup 2, agrees with an activation of the transcriptional machinery of cytotoxic cells in subgroup 1 [52]. Our GSEA analysis confirmed these results, and also suggested a likely involvement of B-cells. The favorable prognostic impact of the immune system has been suggested by other studies. In colon cancer, increased levels of mRNA for products of T_H1 cells are associated with prolonged survival [53]. The four prognostic

expression studies dedicated to ER- breast cancers [19-22], as well as a meta-analysis [10], revealed that immune response favorably impacted the clinical outcome. Similar observations were recently reported in highly proliferative tumors that included a majority of ER- samples [54]. Our data also agree with the favorable contribution of the immune response to response to chemotherapy recently reported in breast cancer [19,55,56]. Altogether, these observations suggest a potential interest for therapeutic strategies aimed at stimulating the immune defenses in basal breast cancer. Despite this enrichment for genes involved in immunity, we showed that our model differed from two previously published prognostic immune signatures [19-21]. Indeed, beside the immune system, other overrepresented Ingenuity pathways involved PKR, PPAR and RXR. PKR is a P53 target protein kinase, which plays a crucial role in the tumor-suppressor function of P53 and apoptosis [57]. PPARgamma is a ligand-activated transcription factor that regulates cell proliferation and differentiation. PPARgamma ligands, through the downregulation of gelatinases, inhibit the invasive capacities of breast cancer cells *in vitro* [58], and repress TGFbeta signaling involved in metastasis [59]. Finally, bexarotene (Targretin), an RXR (retinoid X receptor) agonist, inhibits angiogenesis and metastasis *in vitro* after activation of its heterodimerization partner PPARgamma [60].

Regarding the 166 genes overexpressed in the poor-prognosis subgroup 2, we had previously identified [28] several Onto-Express biological processes related to cytoskeleton, muscle biology, adhesion and tyrosine kinase signaling. Similarly, Ingenuity analysis (Supplementary Table 5) identified several pathways related to cell migration such as “caveolar-mediated endocytosis”, “virus entry via endocytic pathways”, “tight junction signaling”, “agrin interactions at neuromuscular junction”, “actin cytoskeleton signaling”, and “clathrin-mediated endocytosis” [61,62]. Indeed,

many genes are involved in the architecture and remodeling of cytoskeleton and adhesion. Examples include *ACTA2*, *ACTG2*, *FLNA*, *FLNC*, *ACTN1*, *MYL9*, *TPM2*, *MYLK*, *M-RIP*, *CALD1*, *CNN2*, *TAGLN*, *DAAM1*, *MTSS1*, *SMTN*, *PARVA*, *FLH1*, and *ADAM12*. Other pathways potentially related to tumor aggressiveness included "VEGF signaling, "G-protein coupled receptor signaling", or "NF-$_k$B activation by viruses". Finally, the presence of genes coding for smooth muscle-specific proteins (*ACTG2*, *ACTA2*, *TPM2*, *MYL9, M-RIP, CALD1*, *CNN2*, *SMTN*, *KCNMB1*, *TAGLN*, *ACTN1*, *APEG1*, *BOC*) and of genes upregulated by TGFbeta [63-65] (*TAGLN*, *ACTG2*, *FHL2, TPM2*, *ACTN1, CNN2, FSTL1, BGN, TGFB1I1*) may suggest some degree of epithelial-to-mesenchymal transition in the poor-prognosis subgroup, which calls for complementary analyses for confirming this hypothesis. . Of note, our analysis did not reveal any differential implication of cell cycle and cell proliferation pathways, in agreement with the high grade of basal tumors, equally distributed between our two prognostic subgroups.

In conclusion, we have shown that early basal breast cancers can be classified in two subgroups with different DFS based on a 368-gene model. This new prognostic classifier was validated in a series of 466 basal breast cancers and outperformed the classical histoclinical features in multivariate analysis. The difference for clinical outcome might be due, at least in part, to an effective host immune T_H1 response, enhanced tumor cell apoptosis, elevated levels of metastasis-inhibiting factors and low levels of metastasis-promoting factors in the good-prognosis subgroup, and a more developed cell migration system in the poor-prognosis subgroup. Clinically, the identification of poor or good prognosis cases within basal breast cancers should help select the appropriate systemic treatment,

while the identification of biologically relevant genes or pathways included in the classifier should provide new potential therapeutical targets. Further validation of our model in a larger retrospective series, then in a prospective series is warranted.

AKNOWLEDGMENTS

Our work is supported by Institut Paoli-Calmettes, Inserm, Institut National du Cancer (Tr 2008), Association pour le Recherche contre le Cancer, Ligue Nationale contre le Cancer (label DB), and Fondation pour la Recherche Médicale (RS 2009).

REFERENCES

1. Bertucci F, Finetti P, Cervera N et al (2006) Gene expression profiling and clinical outcome in breast cancer. Omics 10:429-443

2. Sorlie T, Perou CM, Tibshirani R et al (2001) Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc Natl Acad Sci U S A 98:10869-10874

3. Sorlie T, Tibshirani R, Parker J et al (2003) Repeated observation of breast tumor subtypes in independent gene expression data sets. Proc Natl Acad Sci U S A 100:8418-8423

4. Sorlie T, Wang Y, Xiao C et al (2006) Distinct molecular mechanisms underlying clinically relevant subtypes of breast cancer: gene expression analyses across three different platforms. BMC Genomics 7:127

5. Carey LA, Perou CM, Livasy CA et al (2006) Race, breast cancer subtypes, and survival in the Carolina Breast Cancer Study. JAMA 295:2492-2502

6. Adelaide J, Finetti P, Bekhouche I et al (2007) Integrated profiling of basal and luminal breast cancers. Cancer Res 67:11565-11575

7. Bertucci F, Finetti P, Cervera N et al (2009) How different are luminal A and basal breast cancers? Int J Cancer 124:1338-1348

8. Pusztai LI (2009) Gene expression profiling of breast cancer. Breast Cancer Res 11 Suppl 3:S11
9. Troester MA, Hoadley KA, Sorlie T et al (2004) Cell-type-specific responses to chemotherapeutics in breast cancer. Cancer Res 64:4218-4226.
10. Desmedt C, Haibe-Kains B, Wirapati P et al (2008) Biological processes associated with breast cancer clinical outcome depend on the molecular subtypes. Clin Cancer Res 14:5158-5165
11. van de Vijver MJ, He YD, van't Veer LJ et al (2002) A gene-expression signature as a predictor of survival in breast cancer. N Engl J Med 347:1999-2009.
12. van 't Veer LJ, Dai H, van de Vijver MJ et al (2002) Gene expression profiling predicts clinical outcome of breast cancer. Nature 415:530-536
13. Sotiriou C, Wirapati P, Loi S et al (2006) Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. J Natl Cancer Inst 98:262-272
14. Chang HY, Sneddon JB, Alizadeh AA et al (2004) Gene expression signature of fibroblast serum response predicts human cancer progression: similarities between tumors and wounds. PLoS Biol 2:E7
15. Paik S, Shak S, Tang G et al (2004) A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. N Engl J Med 351:2817-2826
16. Miller LD, Smeds J, George J et al (2005) An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. Proc Natl Acad Sci U S A 102:13550-13555
17. Wang Y, Klijn JG, Zhang Y et al (2005) Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. Lancet 365:671-679.
18. Liu R, Wang X, Chen GY et al (2007) The prognostic role of a gene signature from tumorigenic breast-cancer cells. N Engl J Med 356:217-226
19. Rody A, Holtrich U, Pusztai L et al (2009) T-cell metagene predicts a favorable prognosis in estrogen receptor-negative and HER2-positive breast cancers. Breast Cancer Res 11:R15
20. Teschendorff AE, Caldas CI (2008) A robust classifier of high predictive value to identify good prognosis patients in ER-negative breast cancer. Breast Cancer Res 10:R73

21. Teschendorff AE, Miremadi A, Pinder SE, Ellis IO, Caldas CI (2007) An immune response gene expression module identifies a good prognosis subtype in estrogen receptor negative breast cancer. Genome Biol 8:R157
22. Kreike B, van Kouwenhove M, Horlings H et al (2007) Gene expression profiling and histopathological characterization of triple-negative/basal-like breast carcinomas. Breast Cancer Res 9:R65
23. Da Silva L, Clarke C, Lakhani SRI (2007) Demystifying basal-like breast carcinomas. J Clin Pathol 60:1328-1332
24. Olivotto IA, Bajdik CD, Ravdin PM et al (2005) Population-based validation of the prognostic model ADJUVANT! for early breast cancer. J Clin Oncol 23:2716-2725
25. Dent R, Hanna WM, Trudeau M et al (2009) Time to disease recurrence in basal-type breast cancers: effects of tumor size and lymph node status. Cancer
26. Schneider BP, Winer EP, Foulkes WD et al (2008) Triple-negative breast cancer: risk factors to potential targets. Clin Cancer Res 14:8010-8018
27. Palmieri D, Lockman PR, Thomas FC, al. el (2009) Vorinostat inhibits brain metastatic colonization in a model of triple-negative breast cancer and induces DNA double-strand breaks. Clin Cancer Res 15:6148-6157
28. Bertucci F, Finetti P, Cervera N et al (2006) Gene expression profiling shows medullary breast cancer is a subgroup of basal breast cancers. Cancer Res 66:4636-4644
29. Vincent-Salomon A, Gruel N, Lucchesi C et al (2007) Identification of typical medullary breast carcinoma as a genomic sub-group of basal-like carcinomas, a heterogeneous new molecular entity. Breast Cancer Res 9:R24
30. Weigelt B, Horlings HM, Kreike B et al (2008) Refinement of breast cancer classification by molecular characterization of histological special types. J Pathol 216:141-150
31. Finetti P, Cervera N, Charafe-Jauffret E et al (2008) Sixteen-kinase gene expression identifies luminal breast cancers with poor prognosis. Cancer Res 68:767-776
32. Desmedt C, Piette F, Loi S et al (2007) Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the TRANSBIG multicenter independent validation series. Clin Cancer Res 13:3207-3214

33. Ivshina AV, George J, Senko O et al (2006) Genetic reclassification of histologic grade delineates new clinical subtypes of breast cancer. Cancer Res 66:10292-10301
34. Hess KR, Anderson K, Symmans WF et al (2006) Pharmacogenomic predictor of sensitivity to preoperative chemotherapy with paclitaxel and fluorouracil, doxorubicin, and cyclophosphamide in breast cancer. J Clin Oncol 24:4236-4244
35. Hu Z, Fan C, Oh DS et al (2006) The molecular portraits of breast tumors are conserved across microarray platforms. BMC Genomics 7:96
36. Perreard L, Fan C, Quackenbush JF et al (2006) Classification and risk stratification of invasive breast carcinomas using a real-time quantitative RT-PCR assay. Breast Cancer Res 8:R23
37. Herschkowitz JI, Simin K, Weigman VJ et al (2007) Identification of conserved gene expression features between murine mammary carcinoma models and human breast tumors. Genome Biol 8:R76
38. Hoadley KA, Weigman VJ, Fan C et al (2007) EGFR associated expression profiles vary with breast tumor subtype. BMC Genomics 8:258
39. Mullins M, Perreard L, Quackenbush JF et al (2007) Agreement in breast cancer classification between microarray and quantitative reverse transcription PCR from fresh-frozen and formalin-fixed, paraffin-embedded tissues. Clin Chem 53:1273-1279
40. Oh DS, Troester MA, Usary J et al (2006) Estrogen-regulated genes predict survival in hormone receptor-positive breast cancers. J Clin Oncol 24:1656-1664
41. Parker JS, Mullins M, Cheang MC et al (2009) Supervised risk predictor of breast cancer based on intrinsic subtypes. J Clin Oncol 27:1160-1167
42. Weigelt B, Hu Z, He X et al (2005) Molecular portraits and 70-gene prognosis signature are preserved throughout the metastatic process of breast cancer. Cancer Res 65:9155-9158
43. Irizarry RA, Hobbs B, Collin F et al (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics 4:249-264.
44. Benito M, Parker J, Du Q et al (2004) Adjustment of systematic microarray data biases. Bioinformatics 20:105-114
45. Palmer C, Diehn M, Alizadeh AA, Brown POI (2006) Cell-type specific gene expression profiles of leukocytes in human peripheral blood. BMC Genomics 7:115

46. Subramanian A, Tamayo P, Mootha VK et al (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A 102:15545-15550
47. Kim SYI (2009) Effects of sample size on robustness and prediction accuracy of a prognostic gene signature. BMC Bioinformatics 10:147
48. Sotiriou C, Piccart MJI (2007) Taking gene-expression profiling to the clinic: when will molecular signatures become relevant to patient care? Nat Rev Cancer 7:545-553
49. Chang HY, Nuyten DS, Sneddon JB et al (2005) Robustness, scalability, and integration of a wound-response gene expression signature in predicting breast cancer survival. Proc Natl Acad Sci U S A 102:3738-3743
50. Van Laere S, Beissbarth T, Van der Auwera I et al (2008) Relapse-free survival in breast cancer patients is associated with a gene expression signature characteristic for inflammatory breast cancer. Clin Cancer Res 14:7452-7460
51. Hisada M, Kamiya S, Fujita K et al (2004) Potent antitumor activity of interleukin-27. Cancer Res. 64:1152-1156.
52. Li Q, Verma IMI (2002) NF-kappaB regulation in the immune system. Nat Rev Immunol 2:725-734.
53. Pages F, Berger A, Camus M et al (2005) Effector memory T cells, early metastasis, and survival in colorectal cancer. N Engl J Med 353:2654-2666
54. Schmidt M, Bohm D, von Torne C et al (2008) The humoral immune system has a key prognostic impact in node-negative breast cancer. Cancer Res 68:5405-5413
55. de Kruijf EM, van Nes JG, Sajet A et al (2010) The predictive value of HLA class I tumor cell expression and presence of intratumoral Tregs for chemotherapy in patients with early breast cancer. Clin Cancer Res 16:1272-1280
56. Denkert C, Loibl S, Noske A et al (2010) Tumor-associated lymphocytes as an independent predictor of response to neoadjuvant chemotherapy in breast cancer. J Clin Oncol 28:105-113
57. Yoon CH, Lee ES, Lim DS, Bae YSI (2009) PKR, a p53 target gene, plays a crucial role in the tumor-suppressor function of p53. Proc Natl Acad Sci U S A 106:7852-7857
58. Liu H, Zang C, Fenner MH, Possinger K, Elstner EI (2003) PPARgamma ligands and ATRA inhibit the invasion of human breast cancer cells in vitro. Breast Cancer Res Treat 79:63-74

59. Jarrar MH, Baranova AI (2007) PPARgamma activation by thiazolidinediones (TZDs) may modulate breast carcinoma outcome: the importance of interplay with TGFbeta signalling. J Cell Mol Med 11:71-87
60. Yen WC, Prudente RY, Corpuz MR, Negro-Vilar A, Lamph WWI (2006) A selective retinoid X receptor agonist bexarotene (LGD1069, targretin) inhibits angiogenesis and metastasis in solid tumours. Br J Cancer 94:654-660
61. Chao WT, Kunz JI (2009) Focal adhesion disassembly requires clathrin-dependent endocytosis of integrins. FEBS Lett 583:1337-1343
62. Echarri A, Muriel O, Del Pozo MAI (2007) Intracellular trafficking of raft/caveolae domains: insights from integrin signaling. Semin Cell Dev Biol 18:627-637
63. Bakin AV, Safina A, Rinehart C et aI (2004) A critical role of tropomyosins in TGF-beta regulation of the actin cytoskeleton and cell motility in epithelial cells. Mol Biol Cell. 15:4682-4694.
64. Untergasser G, Gander R, Lilg C et aI (2005) Profiling molecular targets of TGF-beta1 in prostate fibroblast-to-myofibroblast transdifferentiation. Mech Ageing Dev. 126:59-69.
65. Groth S, Schulze M, Kalthoff H, Fandrich F, Ungefroren HI (2005) Adhesion and Rac1-dependent regulation of biglycan gene expression by transforming growth factor-beta. Evidence for oxidative signaling through NADPH oxidase. J Biol Chem. 280:33190-33199.

Table 1. Molecular subtypes and histoclinical correlations

Table 2. Histoclinical characteristics of the two basal tumor subgroups

Table 3. Univariate and multivariate DFS analyses by Cox regression of basal tumors. *A/* without, and *B/* with the immune response (IR) signature-based classification.

Table 4. Prognostic classification of breast cancers using the 368-gene classifier and six prognostic breast cancer signatures.

Supplementary Table 1. Description of the breast cancer data sets

Supplementary table 2: IPC data set histoclinical data

Supplementary Table 3. List of 368 genes differentially expressed between basal MBC and nonMBC.

Supplementary Table 4: Application of six major prognostic breast cancer signatures to our pooled series

Supplementary Table 5. Ingenuity canonical pathways overrepresented in the good-prognosis and the poor-prognosis basal breast cancer subgroups.

Supplementary Figure 1. GSEA shows correlations between our SVM model-based classification of basal breast cancers and cell-type specific gene expression signatures of leucocytes.

A/ Results of GSEA with the five tested signatures. NES, normalized enrichment score; FDR, false discovery rate.

B/ Enrichment plots for the three significant signatures: B-cell, T-cell, and CD8+ T-cell (from left to right).

FIGURE LEGENDS

Fig. 1 Classification of basal MBC and nonMBC based on the 368-gene SVM model

A/ Learning set. SVM model-based classification of 37 IPC basal breast cancers (16 nonMBC, and 21 MBC) from which we had generated the 368-gene signature and defined the SVM model [28]. *Top*: cross-table. *Middle*: Box plots of the SVM prediction score in MBC and nonMBC samples. The dashed horizontal line indicates the threshold 0 that separates two subgroups of samples: subgroup 1 (above the line) and subgroup 2 (under the line). *Bottom*: Classification of samples based on the SVM score. The vertical orange line indicates the threshold 0 that separates the two subgroups of samples (left of the line, subgroup 1; right to the line, subgroup 2). The histological type and the SVM score are color-coded as indicated. *B/* Validation set. The legend is similar to A, but applies to the validation set (basal samples from [30]).

Fig. 2 Disease-free survival of the two basal breast cancer subgroups

A/ The 368-gene SVM model was applied to 466 basal breast cancers and defined two subgroups 1 and 2. *B/* Kaplan-Meier disease-free survival curves of the two basal breast cancer subgroups defined in A.

Table 1: Molecular subtypes and histoclinical correlations

Characteristics (N)	Basal N=466	ERBB2 N=280	Luminal A N=653	Luminal B N=359	Normal-like N=276	*p*-value
	N (% of evaluated cases)					
Age (1661)						3E-05
≤ 50 years	215 (57%)	139 (58%)	231 (43%)	149 (52%)	99 (46%)	
> 50 years	163 (43%)	99 (42%)	312 (57%)	137 (48%)	117 (54%)	
Histological type (558)						1.46E-02
ductal	135 (92%)	49 (92%)	149 (79%)	96 (85%)	41 (72%)	
lobular	4 (3%)	0 (0%)	17 (9%)	6 (5%)	7 (12%)	
mixt	3 (2%)	1 (2%)	13 (7%)	6 (5%)	5 (9%)	
other*	5 (3%)	3 (6%)	9 (5%)	5 (4%)	4 (7%)	
Pathological tumor size pT (1622)						1.00E-05
pT1	105 (28%)	90 (38%)	231 (43%)	90 (32%)	98 (49%)	
pT2-4	265 (72%)	144 (62%)	306 (57%)	190 (68%)	103 (51%)	
Pathological lymph node status pN (1664)						2.77E-03
negative	260 (65%)	99 (51%)	346 (62%)	173 (61%)	156 (68%)	
positive	138 (35%)	96 (49%)	213 (38%)	111 (39%)	72 (32%)	
Tumor grade (1658)						1.00E-05
SBR 1	14 (4%)	16 (7%)	148 (28%)	30 (10%)	69 (32%)	
SBR 2-3	364 (96%)	226 (93%)	389 (72%)	258 (90%)	144 (68%)	
ER IHC status (1923)						1.00E-05
negative	340 (78%)	110 (42%)	68 (11%)	25 (7%)	71 (28%)	
positive	100 (22%)	155 (58%)	556 (89%)	319 (93%)	179 (72%)	
PR IHC status (657)						1.00E-05
negative	132 (85%)	47 (64%)	42 (18%)	30 (26%)	28 (33%)	
positive	24 (15%)	27 (36%)	187 (82%)	84 (74%)	56 (67%)	
ERBB2 IHC status (375)						1.00E-05
negative	89 (86%)	14 (40%)	119 (94%)	66 (82%)	26 (84%)	
positive	14 (14%)	21 (60%)	7 (6%)	14 (18%)	5 (16%)	
P53 IHC status (194)						1.00E-05
negative	20 (40%)	4 (27%)	61 (82%)	24 (71%)	16 (76%)	
positive	30 (60%)	11 (73%)	13 (18%)	10 (29%)	5 (24%)	
KI67 IHC status (202)						1.00E-05
negative	5 (8%)	2 (11%)	40 (57%)	5 (14%)	6 (35%)	
positive	55 (92%)	16 (89%)	30 (43%)	32 (86%)	11 (65%)	
Follow-up, months (1752)						0.1374**
median	81	77	89	91	95	
Disease-free survival (1752)						6.91E-13
5-year DFS	60% (392)	60% (243)	80% (578)	57% (304)	72% (235)	

N, number of tumor samples - out of the 2034 samples - with available information for the corresponding characteristic
*, other types include tubular (n=12), metaplastic (n=6), mucinous (n=5), apocrine (n=1), histiocytoid (n=1) and unknown (n=1)
**, ANOVA test

Table 2. Histoclinical characteristics of the two basal tumor subgroups

Characteristics (N)	Subgroup 1 N=217	Subgroup 2 N=249	*p*-value
	N (% of evaluated cases)		
Age (378)			4.95E-03
≤ 50 years	115 (65%)	100 (50%)	
> 50 years	63 (35%)	100 (50%)	
Pathological tumor size pT (370)			0.25

pT1	44 (25%)	61 (31%)	
pT2-4	130 (75%)	135 (69%)	
Pathological lymph node status pN (398)			0.25
negative	129 (68%)	131 (63%)	
positive	60 (32%)	78 (37%)	
Tumor grade (378)			0.43
SBR 1	5 (3%)	9 (4%)	
SBR 2-3	172 (97%)	192 (96%)	
ER IHC status (440)			9.14E-04
negative	181 (84%)	159 (71%)	
positive	34 (16%)	66 (29%)	
PR IHC status (156)			0.82
negative	77 (86%)	55 (83%)	
positive	13 (14%)	11 (17%)	
ERBB2 IHC status (103)			1.00
negative	58 (87%)	31 (86%)	
positive	9 (13%)	5 (14%)	
P53 IHC status (50)			0.39
negative	13 (46%)	7 (32%)	
positive	15 (54%)	15 (68%)	
KI67 IHC status (60)			0.19
negative	1 (3%)	4 (14%)	
positive	30 (97%)	25 (86%)	
Follow-up, months (392)			1.18E-02
median	94	62	
Disease-free survival (392)			9.93E-05
5-year DFS	71% (177)	50% (215)	

Table 3A: Univariate and multivariate DFS analyses by Cox regression of basal tumors

	Univariate Analysis				
	N	Hazard Ratio	95% CI	*p*-value	
Age > 50 years (vs ≤ 50 years)	323	1.04	0.73-1.47	0.83	
ER IHC status positive (vs negative)	385	0.75	0.51-1.11	0.15	
PR IHC status positive (vs negative)	106	0.68	0.24-1.97	0.48	
Pathological tumor size pT2-4 (vs pT1)	316	1.38	0.92-2.08	0.12	
Pathological lymph node status positive (vs negative)	343	1.6	1.12-2.28	0.0095	3
ERBB2 IHC status positive (vs negative)	69	1.19	0.28-5.13	0.82	
Tumor grade SBR 2-3 (vs SBR 1)	318	2.04	0.65-6.42	0.22	
KI67 IHC status positive (vs negative)	60	0.46	0.10-2.00	0.30	
SVM classifier-based subgroup 1 (vs sugroup 2)	392	0.53	0.38-0.73	0.000127	3

N is the number of patients with data available regarding the analyzed variables

Table 3B: Similar, but including two immune signature-based classifications (Immune Response IR and T-cell

	Univariate Analysis				
	N	Hazard Ratio	95% CI	*p*-value	
Age > 50 years (vs ≤ 50 years)	323	1.04	0.73-1.47	0.83	

ER IHC status positive (vs negative)	385	0.75	0.51-1.11	0.15	
PR IHC status positive (vs negative)	106	0.68	0.24-1.97	0.48	
Pathological tumor size pT2-4 (vs pT1)	316	1.38	0.92-2.08	0.12	
Pathological lymph node status positive (vs negative)	343	1.6	1.12-2.28	0.0095	3
ERBB2 IHC status positive (vs negative)	69	1.19	0.28-5.13	0.82	
Tumor grade SBR 2-3 (vs SBR 1)	318	2.04	0.65-6.42	0.22	
KI67 IHC status positive (vs negative)	60	0.46	0.10-2.00	0.30	
IR signature Good (vs Poor)	392	0.76	0.56-1.04	0.0913	
T-cell metagene Good (vs Poor)	392	0.6	0.43-0.84	0.0031	3
SVM classifier-based subgroup 1 (vs sugroup 2)	392	0.53	0.38-0.73	0.000127	3

N is the number of patients with data available regarding the analyzed variables

Table 4: Prognostic classification of breast cancers using the 368-gene classifier and six pro

Pronostic signatures		All breast cancers			Basal	
		N	HR* [95% CI]	*p*-value	N	HR* [95% CI]
MBC signature	Subgroup 2 vs 1	1752	1.00 [0.83-1.20]	0.98	392	1.89 [1.37-2.63]
70-gene signature	Poor vs Good	1752	2.36 [1.90-2.93]	6.7E-15	392	NaN
Genomic grade index	Poor vs Good	1752	2.15 [1.82-2.54]	<1E-16	392	0.55 [0.23-1.35]
76-gene signature	Poor vs Good	1728	1.78 [1.51-2.09]	2E-12	387	1.37 [0.98-1.92]
Recurrence score	Intermediate vs Good	1752	1.76 [1.40-2.21]	1.20E-06	392	Inf [0-Inf]
	Poor vs Good		2.22 [1.85-2.67]	<1E-16		Inf [0-Inf]
Wound response signature	Poor vs Good	1752	1.91 [1.55-2.35]	1.4E-09	392	1.33 [0.19-9.50]
Invasiveness gene signature	Poor vs Good	1752	1.64 [1.40-1.91]	6.9E-10	392	0.76 [0.52-1.11]

Pronostic signatures		Luminal A			Luminal B	
		N	HR* [95% CI]	*p*-value	N	HR* [95% CI]
MBC signature	Subgroup 2 vs 1	578	0.89 [0.44-1.82]	0.75	304	0.89 [0.60-1.35]
70-gene signature	Poor vs Good	578	1.54 [1.11-2.13]	9.50E-03	304	2.98 [1.39-6.36]
Genomic grade index	Poor vs Good	578	2.64 [1.85-3.78]	1.10E-07	304	2.21 [1.39-3.52]
76-gene signature	Poor vs Good	570	1.44 [1.03-2.00]	3.20E-02	303	1.83 [1.05-3.17]
Recurrence score	Intermediate vs Good	578	1.74 [1.19-2.54]	4.00E-03	304	1.16 [0.75-1.77]
	Poor vs Good		2.02 [1.16-3.51]	1.30E-02		1.79 [1.21-2.65]
Wound response signature	Poor vs Good	578	1.49 [1.06-2.09]	2.10E-02	304	1.18 [0.52-2.66]
Invasiveness gene signature	Poor vs Good	578	1.69 [1.16-2.46]	5.90E-03	304	1.41 [1.00-1.99]

*HR, hazard ratio

Figure 1

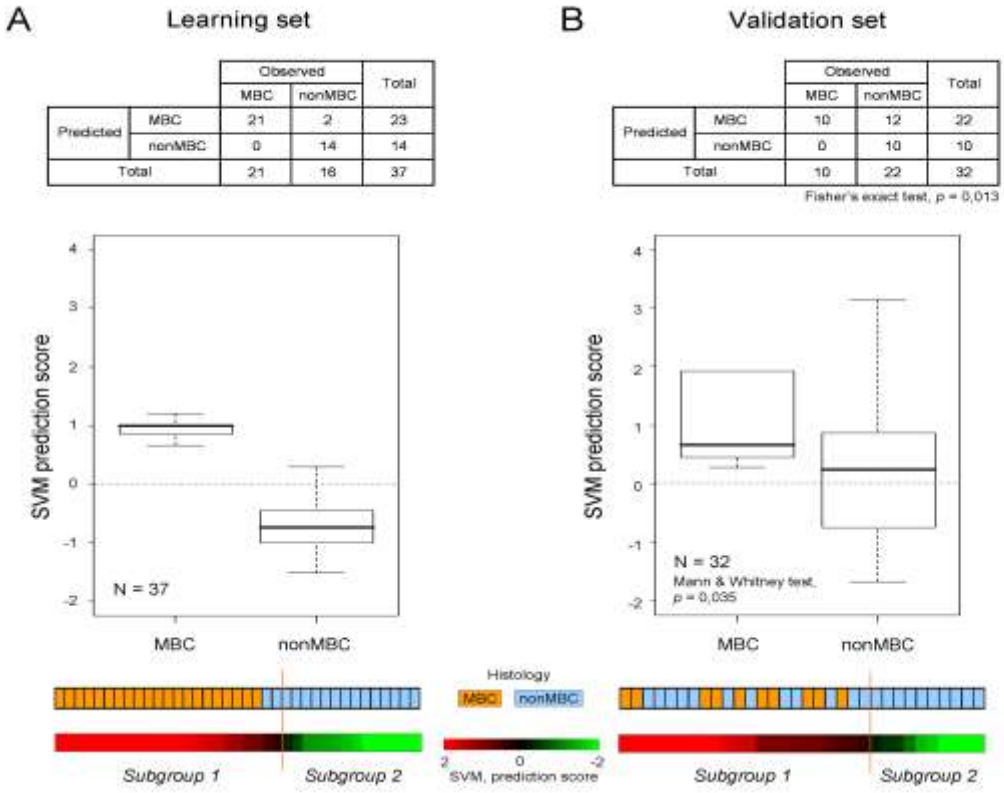

Figure 2